Bioinformatics, 38(21), 2022, 4885–4892
https://doi.org/10.1093/bioinformatics/btac617
Advance Access Publication Date: 9 September 2022
Original Paper

Gene expression

A machine learning-based method for automatically identifying novel cells in annotating single-cell RNA-seq data

Ziyi Li [1,*], Yizhuo Wang [1], Irene Ganan-Gomez[2], Simona Colla[2] and Kim-Anh Do[1,*]

[1]Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, TX, 77030, USA and [2]Department of Leukemia, The University of Texas MD Anderson Cancer Center, Houston, TX, 77030, USA

*To whom correspondence should be addressed.

Associate Editor: Anthony Mathelier

Received on December 7, 2021; revised on September 6, 2022; editorial decision on September 7, 2022; accepted on September 8, 2022

Abstract

Motivation: Single-cell RNA sequencing (scRNA-seq) has been widely used to decompose complex tissues into functionally distinct cell types. The first and usually the most important step of scRNA-seq data analysis is to accurately annotate the cell labels. In recent years, many supervised annotation methods have been developed and shown to be more convenient and accurate than unsupervised cell clustering. One challenge faced by all the supervised annotation methods is the identification of the novel cell type, which is defined as the cell type that is not present in the training data, only exists in the testing data. Existing methods usually label the cells simply based on the correlation coefficients or confidence scores, which sometimes results in an excessive number of unlabeled cells.

Results: We developed a straightforward yet effective method combining autoencoder with iterative feature selection to automatically identify novel cells from scRNA-seq data. Our method trains an autoencoder with the labeled training data and applies the autoencoder to the testing data to obtain reconstruction errors. By iteratively selecting features that demonstrate a bi-modal pattern and reclustering the cells using the selected feature, our method can accurately identify novel cells that are not present in the training data. We further combined this approach with a support vector machine to provide a complete solution for annotating the full range of cell types. Extensive numerical experiments using five real scRNA-seq datasets demonstrated favorable performance of the proposed method over existing methods serving similar purposes.

Availability and implementation: Our R software package CAMLU is publicly available through the Zenodo repository (https://doi.org/10.5281/zenodo.7054422) or GitHub repository (https://github.com/ziyili20/CAMLU).

Contact: zli16@mdanderson.org or kimdo@mdanderson.org

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

The emergence of single-cell RNA sequencing (scRNA-seq) enabled researchers to investigate the cellular compositions and the transcriptomic profiles of human tissues with an unprecedented precision and accuracy (Brouzes *et al.*, 2009). This technology also allows researchers and clinicians to investigate cellular composition changes, identify cell type-specific differential genes and quantify within and across cell type heterogeneity in many human diseases (De Micheli *et al.*, 2020; Soneson and Robinson, 2018). In the past decade, scRNA-seq has been applied to the research of Alzheimer's disease (Mathys *et al.*, 2019), colorectal cancer (Li *et al.*, 2017), autism (Velmeshev *et al.*, 2019), glioblastoma (Patel *et al.*, 2014), leukemia (Petti *et al.*, 2018) and many more. Given the raw data from a scRNA-seq experiment, after pre-processing and quality control steps, a typical analysis pipeline starts with clustering and labeling the different cell types (Luecken and Theis, 2019). Then, based on the research objectives, researchers will perform other downstream analyses such as differential analysis (Soneson and Robinson, 2018), trajectory inference (Herring *et al.*, 2018), compositional analysis (Haber *et al.*, 2017) and cell–cell communication inference (Jin *et al.*, 2021).

As the first step, annotating cells and assigning cell type labels is one of the most important steps since most of the downstream

analyses rely on the accuracy of the cell labels. The traditional way of annotating cells is to apply unsupervised clustering and label the cell types based on the cluster-specific markers (Duò *et al.*, 2018). A series of mature methods have been developed to cluster cells using a variety of techniques, including k-means (Grün *et al.*, 2015; Kiselev *et al.*, 2017), hierarchical clustering (Lin *et al.*, 2017; Yau *et al.*, 2016), community detection (Satija *et al.*, 2015; Wolf *et al.*, 2018) and model-based approach (Ji and Ji, 2016). Although the clustering methods are well developed, the marker-based annotation often poses great challenges in the data analysis. The annotation procedure is labor intensive and time consuming (Clarke *et al.*, 2021). Moreover, the reproducibility of the annotation results usually cannot be guaranteed due to the variation in the understanding of cell type markers among different researchers.

Recently, with the rapid collection of labeled scRNA-seq data, supervised cell annotation methods have been developed to quickly and reproducibly assign cell labels for the new datasets (Abdelaal *et al.*, 2019). These methods can be grouped into two categories depending on the information required. The first group of methods needs the input of cell type marker information, examples include DigitalCellSorter (Domanskyi *et al.*, 2019), Garnett (Pliner *et al.*, 2019), CellAssign (Zhang *et al.*, 2019a), SCINA (Zhang *et al.*, 2019c) and others. The second group of methods requires prelabeled scRNA-seq to train a machine learning classifier and apply this classifier to the new data to get labels, such as scmap (Kiselev *et al.*, 2018), CHETAH (de Kanter *et al.*, 2019), scPred (Alquicira-Hernandez *et al.*, 2019) and others. Compared with unsupervised methods, supervised approaches are generally faster and more convenient. The review paper by Abdelaal *et al.* (2019) compared a total of 26 supervised methods and concluded that support vector machine, scmap-cluster and scPred are among the top, most accurate methods of all the existing methods for annotating cells. Some recent review papers also comprehensively summarized different strategies to establish references for the supervised annotation methods (Ma *et al.*, 2021).

Although multiple supervised methods provide various solutions for annotating cells, one big challenge in all supervised methods is to differentiate the novel (or unknown) cell type from the known cell types. The novel cell type is defined as the cell type that is not present in the training data, only exists in the testing data. Most of the conventional machine learning classification methods can only identify cell types that exist in the training data. When novel cells exist, current methods usually use a cutoff for correlation coefficients between the training and testing data or confidence scores to separate cells with higher confident labels from those with lower confidence (de Kanter *et al.*, 2019; Kiselev *et al.*, 2018). The cells with lower confidence or correlations are usually annotated as 'unassigned'. However, such unassigned cells can be a mixture of novel cells, which are cell types not presenting in the training data, and uncertain cells, which are cell types included in the training data that are difficult to assign due to the high similarities between cell types. Such an unknown-cell-identification strategy may not be ideal, as it usually leads to an excessive number of unlabeled cells.

In this work, we developed a new two-step approach to automatically label scRNA-seq data that contain novel cells. We call it Cell Annotation using a Machine Learning-based method for the presence of Unknown cells (CAMLU). CAMLU was inspired by the applications of autoencoder to detect data outliers in other fields such as computer science (Kieu *et al.*, 2019; Wan *et al.*, 2019) and finance (Demestichas *et al.*, 2021). In the first step, CAMLU uses a combination of autoencoder and iterative feature selection to distinguish known cell types from the novel cell type. The intuition behind this step is that, after training the autoencoder with the training data, the autoencoder will contain the information from all the known cell types. Applying this autoencoder to the testing data will generate reconstruction errors for all the genes. As the cells are a mixture of known and unknown cell types, a few "informative" genes will have a bi-modal distribution in their reconstruction errors, representing their different levels of similarity with the known cell types. Through iterative feature selection, CAMLU can select a smaller set of informative features that have expression differences in the known and unknown cell populations, and finally distinguish novel cells from known cell types. We then can re-cluster the cells based on these informative genes and identify the novel cells. Removing the novel cells identified in the first step, CAMLU uses a support vector machine to exhaustively annotate the rest of the cells in the second step. In the following sections, we first present the technical details of CAMLU. We then evaluate the proposal through a series of extensive numerical experiments using five real datasets. Compared with existing supervised methods, CAMLU demonstrates favorable accuracy in identifying the novel cells and in annotating all the cell types.

2 Materials and methods

Figure 1 illustrates the schematic pipeline of how CAMLU works. Denote the scRNA-seq expression matrix of the training data by $\boldsymbol{Y}_0$ and the corresponding cell labels as $\boldsymbol{B}_{\text{train}}$, where $\boldsymbol{Y}_0$ is a p by n_0 matrix with p being the total number of measured genes and n_0 is the number of cells, $\boldsymbol{B}_{\text{train}}$ is an n_0 by 1 vector. Similarly, denote the testing data by $\boldsymbol{Y}_1$, which has dimensions p by n_1. The first step of CAMLU is to normalize $\boldsymbol{Y}_0$ and $\boldsymbol{Y}_1$ using scater (McCarthy *et al.*, 2017). We select the top 3000 most variable genes from $\boldsymbol{Y}_0$ and the same set of features from $\boldsymbol{Y}_1$. For simplicity, we still use $\boldsymbol{Y}_0$ and $\boldsymbol{Y}_1$ to represent the normalized training and testing datasets after feature selection. We train an autoencoder model with one input layer, one output layer, and five hidden layers:

$$\begin{aligned}\hat{\boldsymbol{Y}}_0 &= \eta(\boldsymbol{z}_{\text{out}}\boldsymbol{W}_{\text{out}} + \boldsymbol{\beta}_{\text{out}})\\ \boldsymbol{z}_{\text{out}} &= \eta(\boldsymbol{z}_5\boldsymbol{W}_5 + \boldsymbol{\beta}_5)\\ &\cdots\\ \boldsymbol{z}_2 &= \eta(\boldsymbol{z}_1\boldsymbol{W}_1 + \boldsymbol{\beta}_1)\\ \boldsymbol{z}_1 &= \eta(\boldsymbol{Y}_0\boldsymbol{W}_0 + \boldsymbol{\beta}_0).\end{aligned}$$

The parameter set $\boldsymbol{\Theta} = \{\boldsymbol{W}_0, \boldsymbol{W}_1, \ldots, \boldsymbol{W}_5, \boldsymbol{\beta}_0, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_5, \boldsymbol{W}_{\text{out}}, \boldsymbol{\beta}_{\text{out}}\}$ is to be estimated during the training process. And $\boldsymbol{z}_l$, for $l = 1, \ldots, 5$, are the hidden neurons with corresponding weight $\boldsymbol{W}_l$ and bias $\boldsymbol{\beta}_l$. $\eta(\cdot)$ is the activation function, which can be a sigmoid, a rectified linear unit (ReLU), rectifiers, or a hyperbolic tangent. As neural networks based on a ReLU function are generally easier to train and can avoid the vanishing gradient problem during optimization (Eckle and Schmidt-Hieber, 2019), we choose to use ReLU, $\eta_{\text{ReLU}}(z) = \max(z, 0)$, as the activation function in all of our experiments. The model is trained using a stochastic gradient descent-based algorithm with the mean squared error loss function $\mathcal{L}(\boldsymbol{Y}_0, \hat{\boldsymbol{Y}}_0) = ||\boldsymbol{Y}_0 - \hat{\boldsymbol{Y}}_0||^2$. We use Adam as the optimization algorithm (Wang *et al.*, 2019) and the mini-batch training strategy (Li *et al.*, 2014), which randomly trains a small proportion of samples in each iteration to improve training efficiency. The number of neurons in the five hidden layers are selected as 256, 128, 64, 128 and 256. We used this structure throughout the experiments and also in our R package implementation.

After the autoencoder has been established, we apply the trained model to the testing data $\boldsymbol{Y}_1$ to obtain the reconstruction $\hat{\boldsymbol{Y}}_1$. The

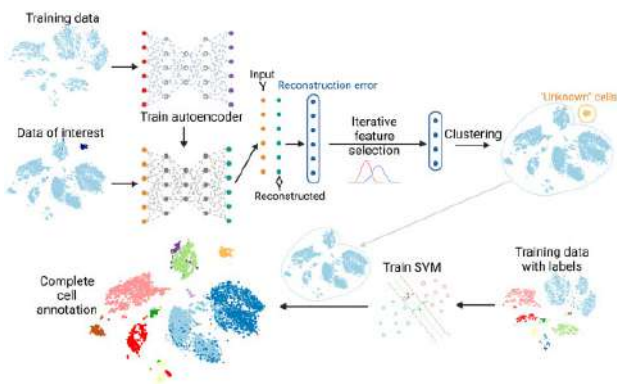

Fig. 1. The working pipeline of CAMLU

reconstruction error is defined as $\mathrm{RE}_{\mathrm{test}} = |\boldsymbol{Y}_1 - \hat{\boldsymbol{Y}}_1|$ and the sum of squared reconstruction errors is $\mathrm{SSE}_{\mathrm{test}} = |\boldsymbol{Y}_1 - \hat{\boldsymbol{Y}}_1|^2$. We design the following iterative feature selection procedure to choose the top informative features for identifying the unknown cells. Denote the clustering results for the testing data at step t by $\boldsymbol{C}_{\mathrm{test}}^{(t)}$. The convergence criterion in Algorithm 1 is defined as the difference between the clustering outcome vectors $\boldsymbol{C}_{\mathrm{test}}^{(t)}$ and $\boldsymbol{C}_{\mathrm{test}}^{(t-1)}$ (i.e. $|\boldsymbol{C}_{\mathrm{test}}^{(t)} - \boldsymbol{C}_{\mathrm{test}}^{(t-1)}|$) smaller than 2 or the number of iterations greater than 10. The cluster with a smaller Pearson's correlation between the mean profile of the cluster-specific cells and the training known cells is defined as the novel cluster. If the algorithm is not converged in 10 iterations or the labels $\boldsymbol{C}_{\mathrm{test}}^{(t)}$ consist of only one group in the middle of the iteration, the major reason is that not enough informative features can be selected. A deeper explanation is that the novel cells (if existing) may be too similar to the known cells. In this situation, the algorithm fails to distinguish them. When this happens, we recommend using traditional biomarker-based annotation to achieve more accurate results. More along this topic is provided in the discussion section.After the unknown cells have been selected, we put them aside from the testing dataset. Denote the labels for the selected novel cells as $\boldsymbol{B}_{\mathrm{test}}^{\mathrm{Novel}}$. We train a support vector machine with linear kernel using the training data $\boldsymbol{Y}_0$ and the labels $\boldsymbol{B}_{\mathrm{train}}$. Then we apply the trained model to the rest of the cells in the testing data to obtain the cell labels $\boldsymbol{B}_{\mathrm{test}}^{\mathrm{Known}}$. The union set $\{\boldsymbol{B}_{\mathrm{test}}^{\mathrm{Novel}}, \boldsymbol{B}_{\mathrm{test}}^{\mathrm{Known}}\}$ contains the full annotation labels from CAMLU. In our current applications, we select the top 3000 variable genes in the initial feature selection step and the top 500 genes in the autoencoder reconstruction error calculation step. Other numbers are evaluated later in simulation studies and the selection have minimum impacts on the detection accuracy as long as a reason number is selected, e.g. more than 3000 genes in the initial selection and top 300–600 genes in the autoencoder reconstruction error calculation step (Supplementary Fig. S14).

Algorithm 1: Iterative feature selection procedure

Data: $\mathrm{RE}_{\mathrm{test}}$ and $\mathrm{SSE}_{\mathrm{test}}$
Result: $\boldsymbol{C}_{\mathrm{test}}$
Initialize $\boldsymbol{C}_{\mathrm{test}}^{(0)}$ by K-means clustering of $\mathrm{SSE}_{\mathrm{test}}$, $K = 2$;
Initialize $t = 1$;
while *Convergence criterion do not meet* **do**
  Perform genewise t test using colttest() function using $\mathrm{RE}_{\mathrm{test}}$ with two groups defined in $\boldsymbol{C}_{\mathrm{test}}^{(t-1)}$;
  Identify the top 500 significant genes based on the testing p-values;
  Update $\boldsymbol{C}_{\mathrm{test}}^{(t)}$ by hierarchical clustering using the selected 500 features, $K = 2$;
end

3 Results

3.1 Monte Carlo numerical experiments

To extensively evaluate the performance of CAMLU, we designed three Monte Carlo numerical experiments based on real datasets. We compared CAMLU with four popular cell annotation methods that are able to identify unknown cells using the 'unassigned' label: CHETAH (de Kanter *et al.*, 2019), scPred (Alquicira-Hernandez *et al.*, 2019), scmap-cluster and scmap-cell (Kiselev *et al.*, 2018). When the unknown cell type is the tumor cell, like in our first experiment, we also compared CAMLU with copyKAT, a method specifically designed to identify malignant cells based on copy number karyotyping. Note that copyKAT only applies when the unknown cells are aneuploid, and thus it is not applicable to the situations when the novel cells are diploid, e.g. designs of our second and third numerical experiments. The evaluation criteria include classification accuracy and adjusted rand index (ARI) (Santos and Embrechts, 2009). For each method, we evaluate both criteria for distinguishing novel cells from known cell types (aAccuracy, ARI) and for assigning the full spectrum of labels (Accuracy-full, ARI-full). For the first two scenarios, we also present the precision and recall for each cell type. An example simulation implementation can be found in the Github repository https://github.com/yizhuo-wang/CAMLU_Simulation.

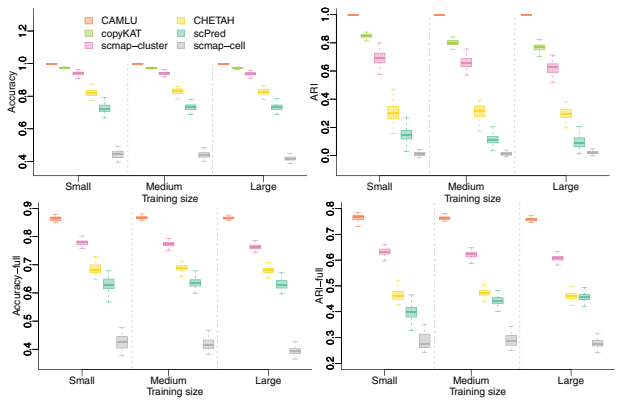

Fig. 2. Results of the numerical experiments using the mixtures of PBMC data and cancer cell line data. The cancer cell line cells are treated as the novel cell type. The upper two panels show the accuracy (left) and ARI (right) of using CAMLU and other existing methods to distinguish known cell types from novel cell types. The lower two panels show the accuracy (left) and ARI (right) of identifying all cell types (CD4T, CD8T, etc.). Results are summarized over 100 Monte Carlo experiments

3.1.1 Numerical experiments with PBMC and HNCC cell line

We obtained the single-cell datasets for peripheral blood mononuclear cells (PBMC) (Zheng *et al.*, 2017) and the head and neck cancer cell line (HNCC) (Kinker *et al.*, 2020). Both data were sequenced by the 10× chromium scRNA-seq technology. The PBMC data has more than 60 000 sorted cells from eight immune cell types. The HNCC has 4632 cancer cells in total. For each experiment, we randomly selected n_1 cells per cell type from the PBMC data and n_2 cancer cells from the HNCC data. We considered three settings with the normal cell sample size $n_1 = 300, 400, 500$ (i.e. 2400, 3100 and 3800 cells in the training data), corresponding to the small, medium and large in Figure 2. The cancer cell number holds constant at $n_2 = 300$ in all settings. In each experiment, we randomly split the selected normal cells into two parts. One part will be used as the training data. The other part will be mixed with the cancer cells and used as the testing data. For all the methods (except copyKAT), we provided the training data to train the classifier and then test the classifier on the testing data for its ability to identify novel cells, as well as assigning other cell labels.

Figure 2 summarized the numerical experiments over 100 Monte Carlo experiments. Compared with existing methods, CAMLU has the highest accuracy in distinguishing the cancer cells from the novel cells and in labeling the full spectrum of cell types. CopyKAT is the second best in identifying cancer cells in the upper panel of Figure 2. Since it cannot assign the full list of labels, copyKAT is not presented in the lower panel of Figure 2. Among other existing methods, scmap-cluster and CHETAH also work well in assigning the correct labels with slightly lower accuracy and ARI, followed by scPred. Scmap-cell has the lowest accuracy for both tasks, probably due to the large number of 'unassigned' labels generated by the method.

The precision and recall results of each cell type were presented in Supplementary Figures S2 and S3. We observed that the proposed method generally has both high precision and recall in either the novel cell identification task or the full cell-type annotation task. Compared to the proposed method, all existing methods have high precision for the known cell types, but low precision in the novel cell type. Scmap-cluster achieves similar recall for both cell types as the

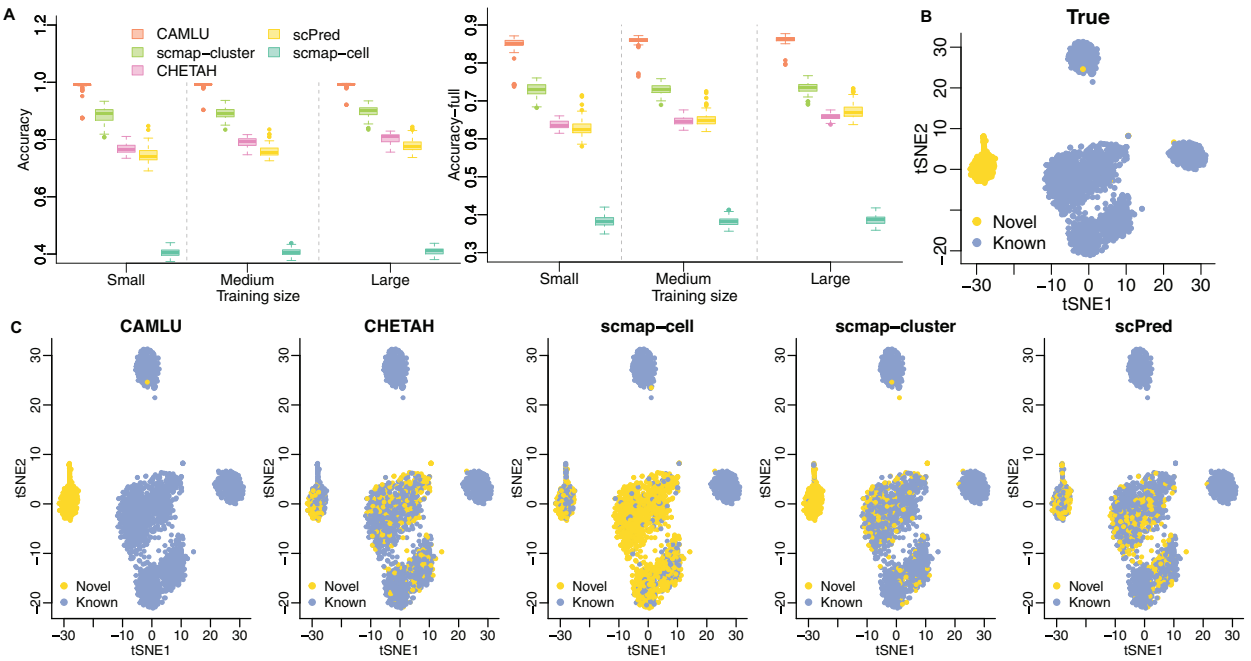

Fig. 3. Results from the numerical experiments using the PBMC data with monocyte as the novel cell type. (**A**) The accuracy of CAMLU and existing methods in distinguishing known cells from novel cells (left) and in identifying all cell types (right). (**B**) The TSNE plot of one simulation experiment the true novel/known status of the cells. (**C**) The identified novel and known cells by CAMLU and other existing methods. Results in (A) are summarized over 100 Monte Carlo experiments

proposed method. All other methods have a lower recall in the known cell type and even lower in the novel cell type. When we look into the full spectrum of cell types, we found that all methods have reasonably well precision and recall in a few cell types (B cells, CD45 NK) but worse results in other cell types, especially the T cell subtypes.

3.1.1 Numerical experiments with PBMC

Next, we designed a numerical experiment with PBMC data only to mimic the situation when the unknown cells are not aneuploid. We treated monocytes as the 'novel' cell type and all other seven cell types as the known cell types. Similar to the first experiment, we randomly selected n_1 cells per cell type for the rest of the seven cell types and n_2 cells from monocytes. We again considered three settings with different sizes of known cell types $n_1 = 300, 400, 500$ and $n_2 = 300$ for monocyte. We randomly split the selected known cells into two parts, one as the training data and the other one mixing with the selected monocytes as the testing data. Since monocytes are not aneuploid, we only compared CAMLU with the four general annotation methods and excluded copyKAT in this comparison.

The accuracies of CAMLU and the existing methods are summarized in Figure 3A. CAMLU has the highest accuracy in identifying the monocytes from the testing data and in assigning all labels. We find scmap-cluster is the second-best method in both tasks, followed by CHETAH and scPred with similar performance. The accuracy by CAMLU drops a little compared to the first numerical experiment, possibly because the current setting is harder. With the increase of training sample size, all methods have slightly improved performance. CAMLU has good accuracy even with the smallest training size. The ARI results in Supplementary Figure S1 have similar conclusions. Figure 3B and C demonstrate the novel cell identification results from a single experiment for true and estimated labels, which may shed light on the differences between CAMLU and the existing method. CAMLU has almost perfect accuracy of distinguishing monocytes from the known cells, while the existing methods, especially scmap-cell, tend to label a lot of known cells as 'unassigned'. We also presented the precision and recall results in Supplementary Figures S4 and S5. Similarly, we observed that the proposed method achieves the highest precision and recall in most of the cell types while existing methods tend to have much lower precision for unknown cell type and lower recall for T cell subtypes.

Using this setting, we also evaluated the impact of fewer novel cells in the testing data on the performance. We considered reasonably large training data (~3000 cells) and a large number of normal cells in the testing data (~3000 cells). We considered the number of unknown cells ranging from smaller numbers (10, 18, 37, 50, 100) to larger numbers (200, 300, 500, 700). The results from 20 Monte Carlo simulations are summarized in Supplementary Figures S10 and S11. We found that the proposed method has comparable performance with existing methods when the novel cell type is very rare, e.g. $< 1\%$, and stays stable when the novel cell type is equal to or more than 1% of the total data. In most of the cases, scmap-cluster and scmap-cell have a similarly stable pattern but with lower accuracies. In contrast, CHETAH and scPred have worse accuracy in all the scenarios.

In real-world scenarios, it is possible that the training data contain a small number of novel or unknown cells. To evaluate the robustness of the proposed method and existing methods, we modified the current simulation setting by including some 'novel' cells in the training dataset and evaluating the annotation accuracy in the testing data. We generated the training and testing data with 500 cells per cell type, and 500 unknown cells in the testing data. Additionally, we added some novel cells ($n = 10, 50, 100, 200$) to the training data to evaluate the robustness of the methods. Supplementary Figure S12 showed the results of including different numbers of unknown cells in the training data. We found that the proposed method is generally robust to including different numbers of novel cells in the training data. Scmap-cluster and scmap-cell also have stable performance when unknown cells are present in the training data. In comparison, both CHETAH and scPred have worse accuracy with the increased number of unknown cells in the training set.

3.1.3 Numerical experiments with pancreas data

In addition to the PBMC data, we also obtained a pancreas scRNA-seq dataset (Muraro *et al.*, 2016) to further evaluate the methods under settings when novel cells are diploid. The dataset was downloaded from Gene Expression Omnibus (GEO) with accession number GSE85241. It contains a total of 2126 single-cell measurements with nine annotated cell types. We designed the experiments by

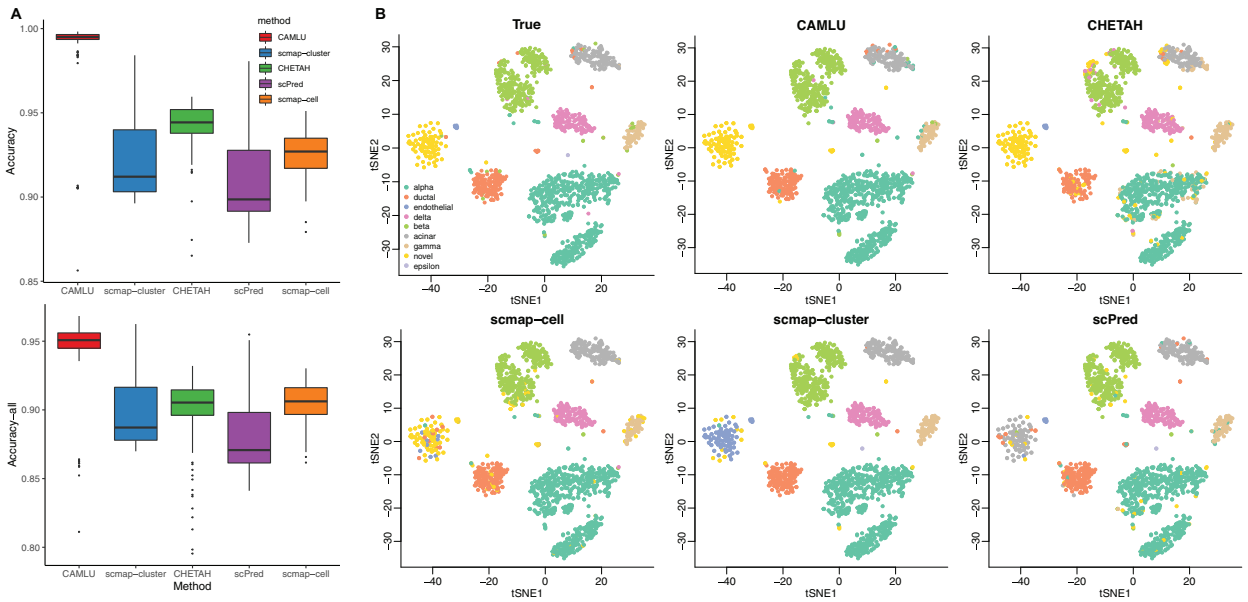

Fig. 4. Results from the numerical experiments using the pancreas data with mesenchymal as the novel cell type. (**A**) The accuracy of CAMLU and existing methods in distinguishing novel cells from known cells (upper) and identifying all cell types (lower). (**B**) The TSNE plot colored with all cell types by true label, CAMLU outputs, and existing method outputs. Results in (A) are summarized over 100 Monte Carlo experiments

treating mesenchymal as the unknown cell type. For each simulation, we randomly drew 500 cells from the 'known' cell types that are not mesenchymal and used the data from these cells as the training data. The rest of the 'known' cells were combined with the mesenchymal cells as the testing data. The number of mesenchymal cells in the data is 80, and the total number of cells in the training data is 1626. This could be a good experiment to evaluate the ability of all the methods in identifying a small number of novel cells.

We summarized the results from 100 Monte Carlo experiments in Figure 4. In Figure 4A, we find that CAMLU has a much higher identification accuracy compared to other existing methods with an average increase of 5–10% in accuracy. In Supplementary Figure S2 and Figure 4B, we visualized the novel cell type and overall annotated cell labels of CAMLU and other methods in comparison to the true labels from one experiment. CAMLU stands out in both tasks and demonstrates high accuracy in distinguishing a very small number of novel cells (<5% of the test data). CHETAH also has good performance in this setting with scmap-cell ranking the third. Both scmap-cluster and scPred fail to identify the novel cells.

3.2 Applications to two real cancer datasets

Lastly, we examined the performance of all the methods in a different setting, cross-subject cell annotation. In all the previous numerical experiments, the cells were from one subject. Or if the cells were from multiple subjects, we mixed the cells and treated them as if they were from the same subject. Due to the big biological variations across different people subject effects sometimes can mix with batch effects and have big impacts on the cell clustering and annotation. In this experiment, we obtained one scRNA-seq dataset including five triple negative breast cancer (TNBC) patients and another dataset including five anaplastic thyroid cancer (ATC) patients. Both the TNBC and ATC datasets were obtained from GEO with accession number GSE148673 (Gao *et al.*, 2021). The normal/malignant labels were provided from the original study (Gao *et al.*, 2021) based on the inferred aneuploid/diploid status of the cells. To evaluate CAMLU and the existing methods, we took out the normal cells from one patient and used those as the training data. We then predicted the normal/malignant status, as well as the full cell labels of the cells from another patient, i.e. the testing data. This experiment design mimics the situation where we use the tumor-adjacent normal tissue from one patient as the training data to predict the cell annotation of the tumor tissue from another patient. The original study (Gao *et al.*, 2021) only provided the normal/malignant status of the cells, not the full annotation. For the cell types other than malignant cells, we used a combination of marker-based annotation (Zhang *et al.*, 2019b) and a reference-based automatic cell annotation methods (Hao *et al.*, 2021) to obtain the benchmark labels.

We first evaluated the accuracy of all the methods in identifying malignant cells from the testing data. In Figure 5B, the cross-subject cell annotation introduces extra noise to the analysis and all the methods have lower accuracies compared with the previous settings. Among all the methods, CAMLU is still the most accurate one for the task, with a mean accuracy around 0.9. The second-best performing method is scmap-cluster in the TNBC dataset and scPred in the ATC dataset. In the TNBC data, scmap-cluster has an accuracy of around 0.80 and all other methods have mean accuracy lower than 0.6. For ATC, all the existing methods have accuracy around0.5–0.6. The advantages of CAMLU also hold if we consider the accuracy in assigning all the cell labels (Supplementary Fig. S3).

As our method assumes that the selected features have different patterns of reconstruction errors in known versus novel cells, it is helpful to examine the distributions of the top features to better understand the proposed methods. In Figure 5B, we illustrated the distribution of the reconstruction errors for four top genes in normal (blue bars) and malignant cells (red bars). For example, COL6A2 encodes one of the three alpha chains of type VI collagen and has been reported to promote tumor progression by affecting both tumor and stromal cells (Chen *et al.*, 2013). We find that COL6A2 has much higher reconstruction error in malignant cells compared with in normal cells, indicating that COL6A2 could be a differential gene between the two cell groups.

Figure 5C shows the most significant Hallmark terms using the top 200 selected features from CAMLU using TNBC and ATC data. We find that a few disease-associated terms have been identified in the results. For example, in TNBC, TNF-α signaling via NF-κB pathway is the most significant Hallmark term. Numerous existing studies have reported the promotion effect of the inflammatory factor TNF-α on the growth of breast cancer (Bauer *et al.*, 2017; Cho *et al.*, 2009). Similarly, the top term in the ATC results, epithelial–mesenchymal transition is an important mechanism related with epithelial tumor progression, local invasion and metastasis. Several studies have reported the strong correlation of epithelial–mesenchymal transition and the progression of ATC (Shakib *et al.*, 2019; Yang *et al.*, 2015).

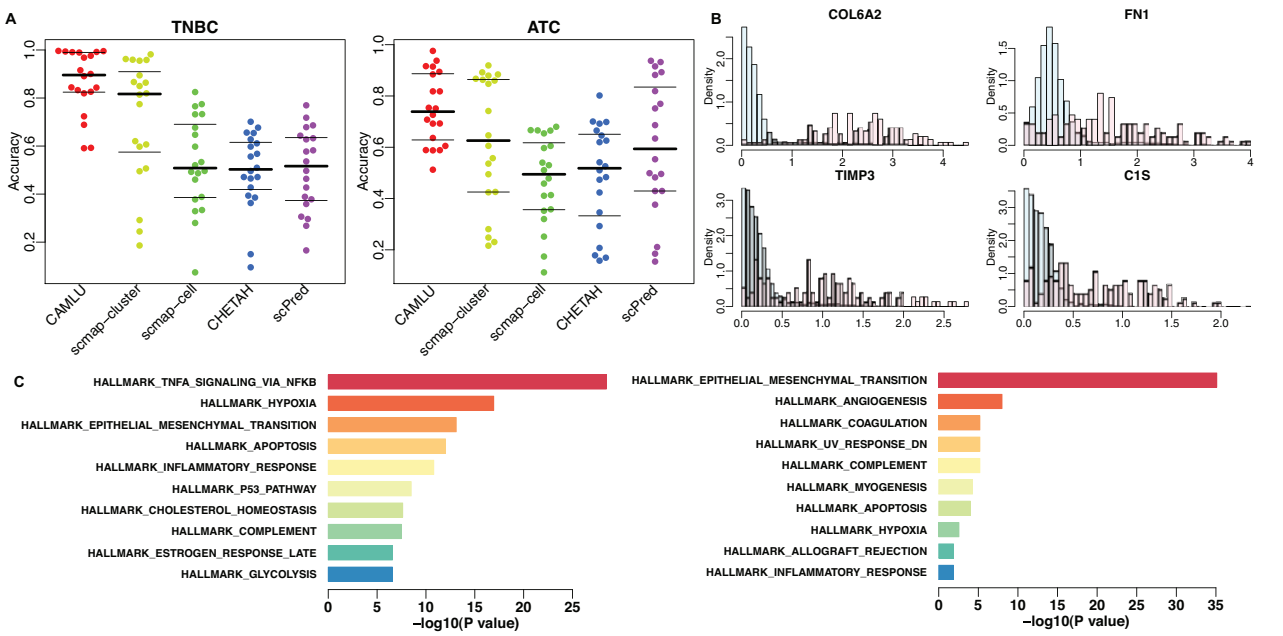

Fig. 5. Results of applying CAMLU and existing methods on two cancer datasets, the triple negative breast cancer (TNBC) and the anaplastic thyroid cancer (ATC). (**A**) The pair-wise accuracy of CAMLU and existing methods in distinguishing malignant cells from known cells. (**B**) The reconstruction error of four genes in malignant cells (lower bars on right) and normal cells (taller bars on left) from CAMLU in the ATC experiment (data from subject GSM4476491 as the training set and GSM4476492 as the testing set). (**C**) The MSigDB Hallmark enrichment analysis using the top 200 genes with differentially reconstructed errors from TNBC and ATC experiments

To evaluate the performance of the proposed method in the scenario when the training and testing data are from different sources, we identified an external TNBC scRNA-seq dataset (GEO accession number 176078; Wu *et al.*, 2021). This study collected the scRNA-seq data from 26 primary tumors of three major breast cancer subtypes, 11 estrogen receptor positive, 6 human epidermal growth factor receptor 2 (HER2) positive and 10 TNBC patients. We took out the data from the 10 TNBC subjects and used the annotated normal cells as the training data to predict the tumor cells in the TNBC scRNA-seq dataset (GSE148673) that analyzed above. Supplementary Figure S8 showed the performance of our proposed method versus other existing methods. We found that even with outside training data, our method still can achieve an accuracy of around 0.8–0.9. We also presented the precision and recall results from this experiment in Supplementary Figure S9. We found that most methods have a reasonable precision in normal cells but very low precision for identifying tumor cells.

4 Discussion

In this work, we developed a novel machine learning-based method for identifying unknown cells from scRNA-seq data. Our pipeline utilizes a combination of autoencoder and iterative feature selection to identify the novel cells based on the reconstruction errors of informative features. After identifying the 'unknown' cells, the rest of the cells are annotated using a support vector machine. Compared with most of the existing methods that label cells with low correlations or low confidence scores as the novel cells, the proposed method innovatively separates the unknown cell selection and the existing cell type annotation. The autoencoder and iterative feature selection method can more accurately identify the novel cells, while the support vector machine-based annotation assigns the cells in an exhaustive way and reduces the number of 'uncertain' cells. As shown in our experiments, the proposed method can achieve higher accuracy in identifying the novel cells in the testing data and then improve the overall assignment accuracy over existing methods.

The proposed method has a few advantages. First, our method does not rely on the aneuploid/diploid cell status. As a result, our method can be applied to identify aneuploid or diploid novel cells, which may imply possibilities of wider application compared to some existing methods. Second, our method can apply to identify different sizes of novel cells. For example, in the pancreas experiment, there are only 80 novel cells out of the more than 1000 cells in the testing data. Our proposed method is able to accurately identify the novel cells while some of the existing methods fail to find the entire cluster. Third, although cross-subject prediction introduces additional noise to the problem and lowers the accuracy in all the methods, our results show that the proposed method still achieves higher performance than existing methods serving similar purposes.

Autoencoder is a widely used machine learning method that has been applied to speech denoising (Lu *et al.*, 2013), representation learning (Tschannen *et al.*, 2018), feature selection (Meng *et al.*, 2017) and many other research areas. Many recent methods for single-cell data analyses are also based on autoencoder for denoising, clustering and other purposes (Lopez *et al.*, 2018; Eraslan *et al.*, 2019; Tran *et al.*, 2021; Lotfollahi *et al.*, 2022). Our proposal is among the first methods applying autoencoder in single-cell data analysis for novel/unknown cell identification. Similar to other machine-learning and deep-learning methods, there are a few parameters that can be selected and tuned to further improve the results, including the number of hidden layers, number of nodes per layer, different normalization methods, different activation function, etc. In the current experiments, we only presented the results using one set of selections and the current results already demonstrate advantages over existing methods. It is possible that the model can be further tuned to improve performance. We will continue to explore other possible model formulation and parameter selections in our future works.

A common concern associated with applying deep learning methods in single-cell data is the computational performance. We comprehensively evaluated the computation performance of the proposed method and existing methods on a desktop with 8-Core Intel i9 processor and 32 Gb memory. The autoencoder was computed on the CPU of the computer and we didn't use GPU for this evaluation. We considered different numbers of cells in the training data (3000, 5000, 8000, 10 000, 15 000 and 20 000) when the testing data are fixed at 1000, and different numbers of cells in the testing data (500, 800, 1000, 1500, 2000 and 3000) when the training data are fixed at 5000. The computational time of our proposed method and other existing methods are presented in the Supplementary Figure S15. We found that the size of the training dataset is closely

associated with the computational time of the proposed method. With a training dataset with less than 10000 cells, the computational cost of CAMLU is quite small ($\leq$5 min). When the number of cells in the training data increases to 20 000, CAMLU takes about 25 min, which is still less than the time required by scPred (>1 h). Overall, the proposed method has a satisfactory computational performance.

The fact that the proposed method does not rely on any biological knowledge is a double-edged sword. On one hand, it is easier to apply and can be used in situations with aneuploid or diploid novel cells. On the other hand, the ability of identifying novel cell type could be limited by the underlying differences between the novel and known cell types. For example, we found that our method may not perform as desired in settings when the novel cells are very similar to the known cell types. For the second numerical experiments using the PBMC data, we tried using B cells as the novel cell type. Compared with monocytes, B cells are more similar to the known cell types (e.g. CD4T and CD8T). The results in Supplementary Figure S13 show that CAMLU has comparable performance to existing methods but larger variation than the results using Monocytes as novel cells (Fig. 3). Based on these observations, we suggest users obtain biological insights before and after applying the proposed method. When the novel cells are highly similar to the normal cells, it may be helpful to incorporate additional biological knowledge (e.g. genotyping and cytogenetics information) in the data analysis (van Galen *et al.*, 2019). It is also important to validate the results using biological biomarkers and interpretations.

In reality, researchers may not know whether the data contain novel cell type or not in many situations. It is helpful to be aware of the method outcome when the data have no novel cells. There are two possible signs for such a scenario, one is failure to converge with the warning 'Cannot detect any novel cells! Returning potential cluster based on the last iteration', and the other one is that a known cell type is picked up by the proposed method. The reported known cell type tends to be the most different one among all the cell types. We evaluated CAMLU in the PBMC-based simulation study with no novel cells in the testing data. Among 20 Monte Carlo iterations, CAMLU reported failure to detect novel cells in 5 datasets and identify CD56NK as the novel cell in the rest of 15 datasets. As the known cell types are already well understood, it is straightforward to identify the true cell identity (i.e. CD56NK) of these 'novel' cells. As a result, we emphasize the importance of validating the results from CAMLU using biological knowledge in collaboration with domain experts.

There are a few directions in which future works can be considered and explored. First, we will continue to explore the selection of different parameters to improve the sensitivity and robustness of the method. For example, different complexity of cell type structure may need larger or smaller autoencoder models. The number of features selected can also be associated with the problem of interest. Adaptive procedures can be designed to automatically select these factors in the model construction. Second, one can consider to better tailor the tool for different disease settings by incorporating additional biological knowledge into the framework. In our current feature selection setting, we select the top features merely based on the reconstruction distributions. It is possible that combining the bimodal features with disease-associated features can achieve even better performance.

Acknowledgements

The authors thank Sunyi Chi from MD Anderson Cancer Center for the helpful discussion at an early stage of the project.

Funding

This work was supported by the National Institutes of Health (NIH) [P30CA016672, P50CA140388, UL1TR003167 and 5R01GM122775], the MD Anderson Moon Shot Programs, and Cancer Prevention & Research Institute of Texas [RP160693] to K.A.D., in part. S.C. is a Scholar of the Leukemia and Lymphoma Society. CPRIT [RP190295 to S.C., I.G.G. and Z.L.], in part; National Institutes of Health (NIH) [R03CA270725 to Z.L. and Y.W.], in part. This work was also supported by philanthropic contributions to MD Anderson's AML and MDS Moon Shot Program.

Conflict of Interest: none declared.

Data availability

The developed method is implemented as an R package and is freely available at https://github.com/ziyili20/CAMLU. An example simulation example can be found at https://github.com/yizhuo-wang/CAMLU_Simulation. All the analyzed data are publicly available and can be downloaded at the 10X website (https://www.10xgenomics.com/resources/datasets?query=&page=1&configure%5Bfacets%5D%5B0%5D=chemistryVersionAndThroughput&configure%5Bfacets%5D%5B1%5D=pipeline.version&configure%5BhitsPerPage%5D=500) and the Gene Expression Omnibus (GSE157220, GSE85241, GSE148673).

References

Abdelaal,T. *et al.* (2019) A comparison of automatic cell identification methods for single-cell RNA sequencing data. *Genome Biol.*, **20**, 1–19.

Alquicira-Hernandez,J. *et al.* (2019) scPred: accurate supervised method for cell-type classification from single-cell RNA-seq data. *Genome Biol.*, **20**, 1–17.

Bauer,D. *et al.* (2017) Apigenin inhibits TNFα/Il-1α-induced CCL2 release through IKBK-epsilon signaling in MDA-MB-231 human breast cancer cells. *PLoS One*, **12**, e0175558.

Brouzes,E. *et al.* (2009) Droplet microfluidic technology for single-cell high--throughput screening. *Proc. Natl. Acad. Sci. USA*, **106**, 14195–14200.

Chen,P. *et al.* (2013) Collagen VI in cancer and its biological mechanisms. *Trends Mol. Med.*, **19**, 410–417.

Cho,S.-G. *et al.* (2009) KiSS1 suppresses TNFα-induced breast cancer cell invasion via an inhibition of RhoA-mediated NF-κB activation. *J. Cell. Biochem.*, **107**, 1139–1149.

Clarke,Z.A. *et al.* (2021) Tutorial: guidelines for annotating single-cell transcriptomic maps using automated and manual methods. *Nat. Protoc.*, **16**, 2749–2764.

de Kanter,J.K. *et al.* (2019) CHETAH: a selective, hierarchical cell type identification method for single-cell RNA sequencing. *Nucleic Acids Res.*, **47**, e95.

De Micheli,A.J. *et al.* (2020) Single-cell transcriptomic analysis identifies extensive heterogeneity in the cellular composition of mouse achilles tendons. *Am. J. Physiol. Cell Physiol.*, **319**, C885–C894.

Demestichas,K. *et al.* (2021) An advanced abnormal behavior detection engine embedding autoencoders for the investigation of financial transactions. *Information*, **12**, 34.

Domanskyi,S. *et al.* (2019) Polled digital cell sorter (p-DCS): automatic identification of hematological cell types from single cell RNA-sequencing clusters. *BMC Bioinformatics*, **20**, 1–16.

Duò,A. *et al.* (2018) A systematic performance evaluation of clustering methods for single-cell RNA-Seq data. *F1000Research*, **7**, 1141.

Eckle,K. and Schmidt-Hieber,J. (2019) A comparison of deep networks with ReLU activation function and linear spline-type methods. *Neural Netw.*, **110**, 232–242.

Eraslan,G. *et al.* (2019) Single-cell RNA-Seq denoising using a deep count autoencoder. *Nat. Commun.*, **10**, 1–14.

Gao,R. *et al.* (2021) Delineating copy number and clonal substructure in human tumors from single-cell transcriptomes. *Nat. Biotechnol.*, **39**, 599–608.

Grün,D. *et al.* (2015) Single-cell messenger RNA sequencing reveals rare intestinal cell types. *Nature*, **525**, 251–255.

Haber,A.L. *et al.* (2017) A single-cell survey of the small intestinal epithelium. *Nature*, **551**, 333–339.

Hao,Y. *et al.* (2021) Integrated analysis of multimodal single-cell data. *Cell*, **184**, 3573–3587.e29.

Herring,C.A. *et al.* (2018) Unsupervised trajectory analysis of single-cell RNA-Seq and imaging data reveals alternative tuft cell origins in the gut. *Cell Syst.*, **6**, 37–51.e9.

Ji,Z. and Ji,H. (2016) TSCAN: pseudo-time reconstruction and evaluation in single-cell RNA-Seq analysis. *Nucleic Acids Res.*, **44**, e117–e117.

Jin,S. *et al.* (2021) Inference and analysis of cell-cell communication using cellchat. *Nat. Commun.*, **12**, 1–20.

Kieu,T. *et al.* (2019) Outlier detection for time series with recurrent autoencoder ensembles. In: *IJCAI*. pp. 2725–2732.
Kinker,G.S. *et al.* (2020) Pan-cancer single-cell RNA-Seq identifies recurring programs of cellular heterogeneity. *Nat. Genet.*, **52**, 1208–1218.
Kiselev,V.Y. *et al.* (2017) SC3: consensus clustering of single-cell RNA-Seq data. *Nat. Methods*, **14**, 483–486.
Kiselev,V.Y. *et al.* (2018) scmap: projection of single-cell RNA-Seq data across data sets. *Nat. Methods*, **15**, 359–362.
Li,H. *et al.* (2017) Reference component analysis of single-cell transcriptomes elucidates cellular heterogeneity in human colorectal tumors. *Nat. Genet.*, **49**, 708–718.
Li,M. *et al.* (2014) Efficient mini-batch training for stochastic optimization. In: *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. pp. 661–670.
Lin,P. *et al.* (2017) CIDR: ultrafast and accurate clustering through imputation for single-cell RNA-Seq data. *Genome Biol.*, **18**, 59–11.
Lopez,R. *et al.* (2018) Deep generative modeling for single-cell transcriptomics. *Nat. Methods*, **15**, 1053–1058.
Lotfollahi,M. *et al.* (2022) Mapping single-cell data to reference atlases by transfer learning. *Nat. Biotechnol.*, **40**, 121–130.
Lu,X. *et al.* (2013) Speech enhancement based on deep denoising autoencoder. In: *Interspeech*, Vol. **2013**. pp. 436–440.
Luecken,M.D. and Theis,F.J. (2019) Current best practices in single-cell RNA-Seq analysis: a tutorial. *Mol. Syst. Biol.*, **15**, e8746.
Ma,W. *et al.* (2021) Evaluation of some aspects in supervised cell type identification for single-cell RNA-Seq: classifier, feature selection, and reference construction. *Genome Biol.*, **22**, 1–23.
Mathys,H. *et al.* (2019) Single-cell transcriptomic analysis of Alzheimer's disease. *Nature*, **570**, 332–337.
McCarthy,D.J. *et al.* (2017) Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-Seq data in R. *Bioinformatics*, **33**, 1179–1186.
Meng,Q. *et al.* (2017) Relational autoencoder for feature extraction. In: *2017 International Joint Conference on Neural Networks (IJCNN)*. IEEE, pp. 364–371.
Muraro,M.J. *et al.* (2016) A single-cell transcriptome atlas of the human pancreas. *Cell Syst.*, **3**, 385–394.e3.
Patel,A.P. *et al.* (2014) Single-cell RNA-Seq highlights intratumoral heterogeneity in primary glioblastoma. *Science*, **344**, 1396–1401.
Petti,A.A. *et al.* (2018) Mutation detection in thousands of acute myeloid leukemia cells using single cell RNA-sequencing. *BioRxiv*, pp. 434746.
Pliner,H.A. *et al.* (2019) Supervised classification enables rapid annotation of cell atlases. *Nat. Methods*, **16**, 983–986.
Santos,J.M. and Embrechts,M. (2009) On the use of the adjusted rand index as a metric for evaluating supervised classification. In: *International Conference on Artificial Neural Networks*. Springer, pp. 175–184.
Satija,R. *et al.* (2015) Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.*, **33**, 495–502.
Shakib,H. *et al.* (2019) Epithelial-to-mesenchymal transition in thyroid cancer: a comprehensive review. *Endocrine*, **66**, 435–455.
Soneson,C. and Robinson,M.D. (2018) Bias, robustness and scalability in single-cell differential expression analysis. *Nat. Methods*, **15**, 255–261.
Tran,D. *et al.* (2021) Fast and precise single-cell data analysis using a hierarchical autoencoder. *Nat. Commun.*, **12**, 1–10.
Tschannen,M. *et al.* (2018) Recent advances in autoencoder-based representation learning. *arXiv preprint arXiv:1812.05069*.
van Galen,P. *et al.* (2019) Single-cell RNA-Seq reveals AML hierarchies relevant to disease progression and immunity. *Cell*, **176**, 1265–1281.e24.
Velmeshev,D. *et al.* (2019) Single-cell genomics identifies cell type–specific molecular changes in autism. *Science*, **364**, 685–689.
Wan,F. *et al.* (2019) Outlier detection for monitoring data using stacked autoencoder. *IEEE Access*, **7**, 173827–173837.
Wang,Y. *et al.* (2019) Assessing optimizer impact on DNN model sensitivity to adversarial examples. *IEEE Access*, **7**, 152766–152776.
Wolf,F.A. *et al.* (2018) SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.*, **19**, 1–5.
Wu,S.Z. *et al.* (2021) A single-cell and spatially resolved atlas of human breast cancers. *Nat. Genet.*, **53**, 1334–1347.
Yang,Y.J. *et al.* (2015) Hypoxia induces epithelial-mesenchymal transition in follicular thyroid cancer: involvement of regulation of twist by hypoxia inducible factor-1α. *Yonsei Med. J.*, **56**, 1503–1514.
Yau,C. *et al.* (2016) pcaReduce: hierarchical clustering of single cell transcriptional profiles. *BMC Bioinformatics*, **17**, 1–11.
Zhang,A.W. *et al.* (2019a) Probabilistic cell-type assignment of single-cell RNA-seq for tumor microenvironment profiling. *Nat. Methods*, **16**, 1007–1015.
Zhang,X. *et al.* (2019b) Cellmarker: a manually curated resource of cell markers in human and mouse. *Nucleic Acids Res.*, **47**, D721–D728.
Zhang,Z. *et al.* (2019c) SCINA: a semi-supervised subtyping algorithm of single cells and bulk samples. *Genes*, **10**, 531.
Zheng,G.X. *et al.* (2017) Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.*, **8**, 14049–14012.